Centre International de Phonétique Appliquée (CIPA) – Universitat Autònoma de Barcelona
Langue(s) & Parole nº 10 ISSN 2684-6691 (version numérisée) ; ISSN 2466-7757 (version papier)

Readability, Fluency and Error identification when using Machine Translation and AI to translate medical research texts from English to European Portuguese

Raquel MOREIRA
https://orcid.org/0000-0002-8728-985X; up201709580@edu.letras.up.pt
Faculty of Arts and Humanities, University of Porto (PORTUGAL)

© author's

Suggested citation: MOREIRA, R. (2025), Readability, fluency and error identification when using machine translation and AI to translate medical research texts from English to European Portuguese, *Langue(s) & Parole,* **10**, 179-200, https://doi.org/10.5565/rev/languesparole.148

Abstract

With the improvements in the algorithms used to develop online machine translation tools and the growing relevance of AI there is a raising concern related to the quality of these tools' outputs. Therefore, we sought to evaluate the readability of the results of machine translation and AI for the translation of medical research texts from English into European Portuguese. A total of 16 scientific papers were collected, and two translations of each source text were obtained, using Google Translate and ChatGPT, (n=32). Several readability formulas and indexes were assessed, as well as text fluency, fidelity and linguistic errors' occurrence. An analysis according to the selected translation tool revealed no significant differences between the results originated from Google Translate and ChatGPT, with regards to the quantitative readability analysis, meaning that their performance was similar despite the perceived fluency and fidelity issues. However, discourse errors were more common in the texts originated from ChatGPT, ($p<0.001$). These results are encouraging and support the use of machine translation tools and AI to translate medical research texts; however, human review and post-editing cannot be ruled out since the algorithms' results largely depend on text complexity.

Keywords: machine translation, artificial intelligence, medical research, readability, translation quality

Résumé

Avec l'amélioration des algorithmes utilisés pour développer des outils de traduction automatique en ligne et l'importance croissante de l'IA, la qualité des résultats de ces outils suscite des inquiétudes croissantes. Par conséquent, nous avons cherché à évaluer la lisibilité des résultats de la traduction automatique et de l'IA pour

la traduction de textes de recherche médicale de l'anglais vers le portugais européen. Au total, 16 articles scientifiques ont été collectés et deux traductions de chaque texte à traduire ont été obtenues à l'aide de Google Translate et de ChatGPT (n = 32). Plusieurs formules et indices de lisibilité ont été évalués, ainsi que la fluidité, la fidélité et la fréquence des erreurs linguistiques. Une analyse selon l'outil de traduction sélectionné n'a révélé aucune différence significative entre les résultats issus de Google Translate et de ChatGPT, en ce qui concerne l'analyse quantitative de la lisibilité, ce qui signifie que leurs performances étaient similaires malgré les problèmes perçus de fluidité et de fidélité. Cependant, les erreurs de discours étaient plus fréquentes dans les textes issus de ChatGPT ($p < 0,001$). Ces résultats sont encourageants et plaident en faveur de l'emploi d'outils de traduction automatique et d'IA pour traduire les textes de recherche médicale. Cependant, la relecture et la post-édition humaines ne peuvent être exclues, car les résultats des algorithmes dépendent largement de la complexité du texte.

Mots-clés : traduction automatique, intelligence artificielle, recherche médicale, lisibilité, qualité de la traduction.

Resumen

Con las mejoras en los algoritmos utilizados para desarrollar herramientas de traducción automática en línea y la creciente relevancia de la IA, existe una creciente preocupación relacionada con la calidad de los resultados de estas herramientas. Así pues, pretendemos evaluar la legibilidad de los resultados de la traducción automática y la IA para la traducción de textos de investigación médica del inglés al portugués europeo. Se recopiló un total de 16 artículos científicos y se obtuvieron dos traducciones de cada texto fuente, utilizando Google Translate y ChatGPT, (n = 32). Se evaluaron varias fórmulas e índices de legibilidad, así como la fluidez del texto, la fidelidad y la frecuencia de errores lingüísticos. El análisis según la herramienta de traducción seleccionada no reveló diferencias significativas entre los resultados originados en Google Translate y ChatGPT, con respecto al análisis de legibilidad cuantitativo, lo que significa que su rendimiento fue similar a pesar de los problemas percibidos de fluidez y fidelidad. Sin embargo, los errores de discurso fueron más comunes en los textos originados en ChatGPT, ($p < 0,001$). Estos resultados son alentadores y respaldan el uso de herramientas de traducción automática y de IA para traducir textos de investigación médica; sin embargo, no se puede descartar la revisión y post-edición humanas, ya que los resultados de los algoritmos dependen en gran medida de la complejidad del texto.

Palabras clave: traducción automática, inteligencia artificial, investigación médica, legibilidad, calidad de la traducción.

Resum

Amb les millores en els algorismes utilitzats per a desenvolupar eines de traducció automàtica en línia i la rellevància creixent de la IA, hi ha una preocupació

creixent relacionada amb la qualitat dels resultats d'aquestes eines. Per tant, es va intentar avaluar la llegibilitat dels resultats de la traducció automàtica i la IA per a la traducció de textos d'investigació mèdica de l'anglès al portuguès europeu. Es van recollir un total de 16 articles científics i es van obtenir dues traduccions de cada text font, mitjançant Google Translate i ChatGPT, (n=32). Es van avaluar diverses fórmules i índexs de llegibilitat, així com la fluïdesa del text, la fidelitat i l'aparició d'errors lingüístics. Una anàlisi segons l'eina de traducció seleccionada no va revelar diferències significatives entre els resultats obtinguts de Google Translate i ChatGPT, pel que fa a l'anàlisi quantitativa de llegibilitat, la qual cosa significa que el seu rendiment va ser similar malgrat els problemes de fluïdesa i fidelitat percebuts. Tanmateix, els errors de discurs eren més freqüents en els textos originats a partir de ChatGPT, ($p<0,001$). Aquests resultats són encoratjadors i donen suport a l'ús d'eines de traducció automàtica i d'IA per traduir textos de investigació mèdica; tanmateix, no es pot descartar la revisió humana i la postedició, ja que els resultats dels algorismes depenen en gran mesura de la complexitat del text.

Paraules clau: traducció automàtica, intel·ligència artificial, investigació mèdica, llegibilitat, qualitat de la traducció.

Introduction

Translation refers to the study of grammatical and lexical structures, as well as communicative situations from a text in a source language, towards a text in the target language. This consists of converting text or speech maintaining the original meaning, as well as the intention and style of the sender (Ikramovna, Lazizovna, 2024, 25; Susanthi, 2023, 22). In turn, machine translation can be defined as the translation of a natural source language into a target language through the use of computer systems and automatic production of text based on the source text (Hutchins, 1992, Baumgartner-Bovier, 2003 cited in Chéragui, 2012, 160). Artificial intelligence is often associated with machine translation, being a branch of computer science that seeks to create software capable of solving problems that, typically, only human intelligence would be able to answer (Kenny, 2022, 34, 35).

Technological, medical and scientific advances, together with the research and dissemination of information they foster, create the need to make information derived from health research accessible through multilingual translations increasingly pressing, promoting medical progress based on peer-to-peer collaboration at a global level that contributes to the establishment of standard procedures (Colina *et al.,* 2017, 267; Hoang, 2021 cited in Ismayilli, 2024, 171, 172).

Translation fidelity is directly related to the maintenance of meaning. This concept is associated with the quality of the translation, since the information and messages from the source text must be conveyed in the target text. There should be no omissions, additions or changes in meaning, and the text should be understandable to the reader (Mcdonald, 2020, 5). Fluency, or intelligibility, is related to the degree of the texts' acceptance by native speakers of the target language. This depends on the degree of adherence to the rules and norms established in the target texts' grammar. The produced text must be acceptable and understandable by the reader, structurally and at the level of the concepts it includes (Mcdonald, 2020, 5). Comprehensibility depends, to a large extent, on human evaluation of the target text and is directly related to the other concepts presented before. The understanding of a text is correlated with the interconnection of the ideas that are to be transmitted, as well as to the quality and fluidity of the text, factors that contribute to the understanding of its meaning by the reader (Smith, Taffler, 1992, 84, 87, 88). Finally, for readability, several definitions have already been proposed (Kamandhari, 2020, 8-11). The main focus of this study is to evaluate the readability of machine translation results, defined as the degree of difficulty through which a text, in this case, a translation, is read and understood. This measure can be assessed through a series of methods that take into account lexical and syntactic complexity, being a research topic that has received increasing attention from the scientific community (Antunes, Lopes, 2019, 149; Mcdonald, 2020, 5).

Thus, the main aim of this study is to evaluate and quantify the readability of machine translation results, using texts written in English, in the context of health research, obtained through Google Translate and ChatGPT. Simultaneously, the texts will be evaluated with regards to their quality based on the degree of fidelity to the source text and their fluency in European Portuguese. Additionally, the identification and classification of the types of errors observed in the target texts will be carried out, as well as the evaluation of the occurrence of errors in each of the tools used and, consequently, the comparison of the performance of both of them in this context.

Methods

This is a comparative study of the performance between a freely available online machine translation tool, Google Translate (https://translate.google.com)

and a chatbot widely used worldwide, ChatGPT (https://chatgpt.com/), which is increasingly influential in translation processes.

The corpus was obtained through MEDLINE and Google Scholar databases. The searches were filtered for texts: (i) published in the last 5 years (since 2020) and (ii) that were available without restrictions, published in open access journals or available through the subscription of the University of Porto.

Table 1 lists and describes the main characteristics of the reference corpus, with the source texts. To these texts are added their translations from Google Translate and ChatGPT, making a total of 48 texts. These translations were obtained between January and February 2025.

Both the texts that constitute the reference corpus and the results of the two tools were compared in terms of their fluency, fidelity and readability, qualitatively, through linguistic analysis, and quantitatively, using absolute and relative counts and calculations using validated formulas for assessing readability, notably: Flesch Reading Ease Score (Costa *et al.,* 2006, 8, 9), Flesch Kincaid Grade Level (Kincaid, 1975), Automated Readability Index (ARI) (Smith, Senter, 1967), Coleman Liau Index (Coleman, Liau, 1975), Gunning Fog Index (Gunning, 1952), FORCAST Formula (Pereira *et al.,* 2023), Powers Sumner Kearl Formula (Pereira *et al.,* 2023), Lix Readability Formula (Pereira *et al.,* 2023), and Rix Readability Grade Level (Pereira *et al.,* 2023). These last four indexes are available through the following link: https://readable.com/readability/readability-formulas/, as stated in (Pereira *et al.,* 2023, 37).

ID	Medical specialty	Study design	Geographical location	Title	(Author, Year)
1	Cardiology	Observational	USA and New Zealand	Diagnostic yield, safety, and outcomes of Head-to-pelvis sudden death CT imaging in post arrest care: The CT FIRST cohort study	(Branch *et al.,* 2023)
2	Cardiology	Randomised clinical trial	Australia	Evaluating a communication aid for return of genetic results in families with hypertrophic cardiomyopathy: A randomized controlled trial	(Burns *et al.,* 2023)
3	Cardiology	Systematic review with meta-analysis	USA	Implementation of appropriate use criteria for cardiology tests and procedures: a systematic review and meta-analysis	(Winchester *et al.,* 2021)

ID	Medical specialty	Study design	Geographical location	Title	(Author, Year)
4	Cardiology	Case report	United Kingdom	Case report: Cytokine therapy and an intracoronary autologous bone marrow-derived cell infusion with Impella support in a patient with dilated cardiomyopathy and a severely reduced ejection fraction	(Hall *et al.,* 2022)
5	Cardiac Surgery	Observational	United Kingdom	Nonanemic Iron Deficiency in Cardiac Surgery: A Retrospective Observational Study	(Horwood *et al.,* 2024)
6	Cardiac Surgery	Randomised clinical trial	USA	Randomized controlled trial of remote ischemic preconditioning in children having cardiac surgery	(Law *et al.,* 2024)
7	Cardiac Surgery	Systematic review with meta-analysis	United Kingdom and Australia	Microplegia in cardiac surgery: Systematic review and meta-analysis	(Owen *et al.,* 2020)
8	Cardiac Surgery	Case report	USA	Case Report: Thrombotic Microangiopathic Anemia After Cardiac Surgery	(Sweeney *et al.,* 2024)
9	Nephrology	Observational	Australia	Tunneled Hemodialysis Catheter Tip Design and Risk of Catheter Dysfunction: An Australian Nationwide Cohort Study	(Lazarus *et al.,* 2024)
10	Nephrology	Randomised clinical trial	United Kingdom	A pragmatic, open-label, randomized controlled trial of Plasma-Lyte-148 versus standard intravenous fluids in children receiving kidney transplants (PLUTO)	(Hayes *et al.,* 2024)
11	Nephrology	Systematic review with meta-analysis	United Kingdom, Netherlands and USA	Symptom burden and health-related quality of life in chronic kidney disease: A global systematic review and meta-analysis	(Fletcher *et al.,* 2022)
12	Nephrology	Case report	USA	A Case of COVID-19-Associated Focal Segmental Glomerulosclerosis	(Thomas *et al.,* 2023)
13	Medical Oncology	Observational	USA	Engagement With Daily Symptom Reporting, Passive Smartphone Sensing, and Wearable Device Data Collection During Chemotherapy: Longitudinal Observational Study	(McClaine *et al.,* 2024)
14	Medical Oncology	Randomised clinical trial	United Kingdom	Dysphagia-optimised intensity-modulated radiotherapy versus standard intensity-modulated radiotherapy in patients with head and neck cancer (DARS): a phase 3, multicentre, randomised, controlled trial	(Nutting *et al.,* 2023)
15	Medical Oncology	Systematic review with meta-analysis	Ireland	Minimally Invasive and Open Gastrectomy for Gastric Cancer: A Systematic Review and Network Meta-Analysis of Randomized Clinical Trials	(Davey *et al.,* 2023)

ID	Medical specialty	Study design	Geographical location	Title	(Author, Year)
16	Medical Oncology	Case report	USA	Use of a topical Janus kinase inhibitor in immune checkpoint inhibitor-induced eczematous reaction: a case report	(Powers *et al.*, 2024)

Table 1. Characteristics of the texts that constitute the reference corpus.

Error quantification will be performed using the taxonomy presented by Costa *et al.* (2015), which is based on the method represented in **Figure 1**.

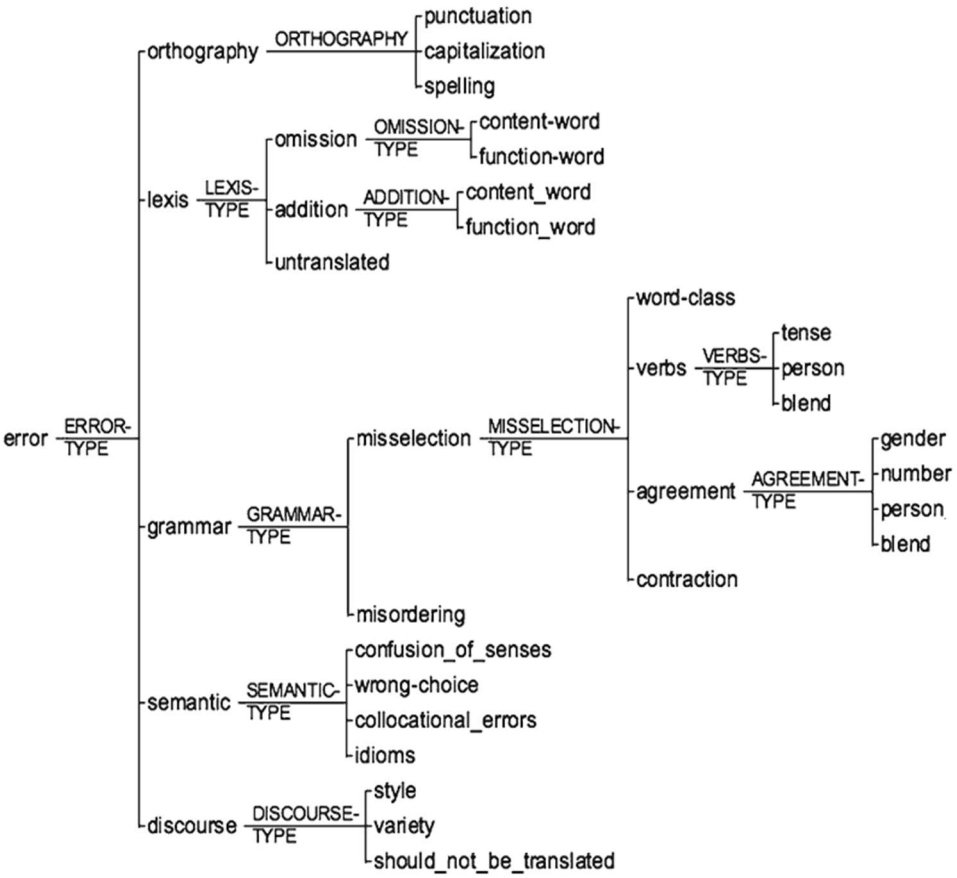

Figure 1. Taxonomy for assessing the quality of machine translation results proposed by Costa et al., *(2015).*

Collins-Thompson (2014) was one of the first authors to propose a methodology for assessing text readability, divided by levels presenting the main aspects related to readability, as shown in **Figure 2**. From this diagram, aspects related to pragmatics, syntax and legibility will be considered for this study since they are not mentioned in the taxonomy presented previously. Errors associated with textual coherence and cohesion will also be considered within the scope of the discourse level assessment given in this pyramid.

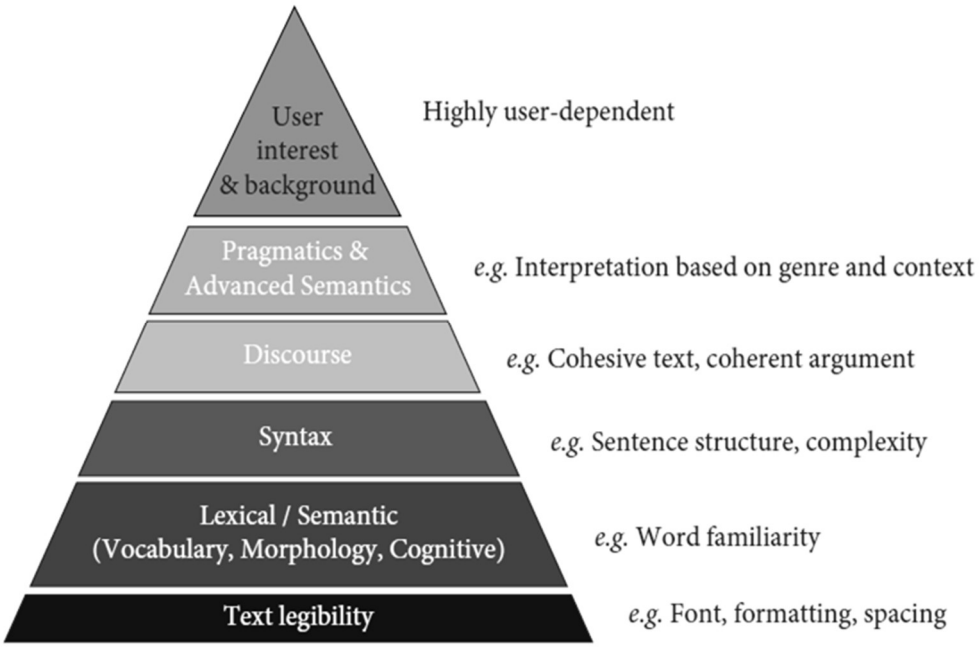

Figure 2. Main aspects to be considered when assessing readability, ordered by levels (Collins-Thompson, 2014, 105).

For the recording and management of variables, as well as for obtaining the quantifications and calculations necessary for the readability assessment, the software Microsoft Word and IBM® SPSS® Statistics (Statistical Package for the Social Sciences), Version 26, were used. To characterise the source and target texts, descriptive statistical measures were obtained, namely, absolute and relative frequencies, as well as central tendency and dispersion measures for the variables classified as continuous on a numerical scale. These procedures allowed us to obtain the readability assessments for each set of texts in the corpus. To study possible associations between variables, inferential statistics methods were used, such as univariate analysis using parametric tests, namely student's t-test for independent samples and ANOVA test or nonparametric Mann-Whitney or Kruskal-Wallis tests, depending on data distribution. For the comparison of proportions between categorical variables and for the comparative statistical analysis between the performance of the two tools (Google Translate and ChatGPT), the chi-square test was used. A significance level of 5% was used. Depending on the distribution of each variable, results are presented in terms of mean ± standard deviation (SD), in the case of these following a normal distribution, or as median and range (minimum – maximum), in the case of variables with a distribution other than normal.

Results

Sixteen original studies were included for analysis, encompassing one text of each of these study designs: case reports, systematic reviews with meta-analysis, observational studies and randomised clinical trials, for each medical specialty selected to create the corpus (cardiology, cardiac surgery, nephrology and medical oncology).

According to the readability analysis carried out by type of tool used to obtain the target texts, no variable revealed statistically significant differences between the results of Google Translate and those of ChatGPT. This reveals that, with regard to quantitative analysis and text readability scores' results, the performance of both tools was similar, despite the recognisable differences in terms of fluency and idiomaticity (detected through human review). Both tools produced texts of similar dimensions and linguistic complexity, using identical sentence structures and the same resources to obtain translations, which is demonstrated through the evaluations presented in **Table 2**.

Variables	Google Translate	ChatGPT	P-value
Text sample, (approximately, 150 words)			
Words, (mean ± SD)	186 ± 15	181 ± 13	0,342
Characters (without spaces), (mean ± SD)	1062 ± 72	1053 ± 76	0,715
Paragraphs, (median (min - max))	2 (1 – 4)	2 (1 – 4)	1,000
Sentences, (median (min - max))	6 (4 – 9)	6 (4 – 9)	0,724
Sentences per paragraph, (median (min - max))	3,3 (2,0 – 7,0)	3,3 (1,7 – 7,0)	0,867
Words per sentence, (median (min - max))	29,8 (19,3 – 48,7)	28,6 (18,1 – 46,2)	0,926
Characters per word, (mean ± SD)	5,6 ± 0,3	5,7 ± 0,2	0,421
Syllables, (median (min - max))	433 (335 – 508)	428 (384 – 516)	0,926
Syllables per word, (mean ± SD)	2,3 ± 0,2	2,4 ± 0,1	0,289
Monosyllables, (median (min - max))	59 (45 – 80)	55 (46 – 77)	0,381
Complex words, (mean ± SD)	77,3 ± 7,6	77,1 ± 8,0	0,946
Complex words (%), (mean ± SD)	41,8 ± 4,1	42,7 ± 4,1	0,519

Readability scores (text sample of approximately, 150 words)			
Flesch Reading Ease Score, (median (min - max))	-24,4 (-43,1 – 29,1)	-26,5 (-45,3 – (-10))	0,361
Flesch-Kincaid Grade Level, (mean ± SD)	25,8 ± 5,8	27,1 ± 5,5	0,543
Automated Readability Index (ARI), (median (min - max))	27,0 (20,0 – 43,9)	27,0 (20,0 – 42,5)	0,809
Coleman Liau Index, (mean ± SD)	20,2 ± 1,1	20,7 ± 1,1	0,169
Gunning Fog Index, (mean ± SD)	320,6 ± 28,0	327,2 ± 26,6	0,500
FORCAST Formula, (median (min - max))	14,1 (12,0 – 15,5)	14,5 (12,3 – 15,4)	0,381
Powers Sumner Kearl Formula, (mean ± SD)	24,8 ± 1,9	24,9 ± 1,8	0,858
Lix Readability Formula, (median (min - max))	30,2 (19,8 – 49,1)	29,0 (18,6 – 46,6)	0,926
Rix Readability Grade Level, (mean ± SD)	12,5 ± 3,0	12,8 ± 3,0	0,760

Footnotes: SD – Standard deviation.

Table 2. Characterization of the texts obtained through Google Translate and ChatGPT.

Table 3 presents the quantifications related to each type of error studied in the corpus composed of the target texts and the differences observed between the two tools used to obtain the translations.

Regarding the types of errors that were significantly different between the two translation tools, discursive errors (which include issues associated with the style of this text genre, the Portuguese variety used and expressions that should not be translated) were substantially more frequent in the target texts obtained through ChatGPT, ((median (min - max)): 46 (9 – 104) vs. 14 (2 – 30), $p < 0.001$). This result was largely due to the recording of multiple errors in the "variety" category in the texts generated by ChatGPT, reflecting the inability of this artificial intelligence tool to differentiate between the European and Brazilian varieties of Portuguese, despite the fact that the desired variety was explicitly mentioned in the interrogative prompt.

Legibility errors, related to text formatting and spacing, were significantly more frequent in texts from Google Translate, ((median (min - max)): 4 (0 – 18) vs. 1 (0 – 8), $p = 0.039$). These were mainly associated

with the inappropriate addition of spaces, while in the results from ChatGPT the most frequent legibility error was the change of the text formatting, for example, with the incorrect presence of bold or the absence of italics, when deemed necessary.

Error type, (median (min - max))	**TOTAL Target Corpus**	**Google Translate**	**ChatGPT**	**P-value**
Full text				
Orthography	2 (0 – 16)	3 (0 – 16)	2 (0 – 4)	0,128
Punctuation	1 (0 – 15)	2 (0 – 15)	1 (0 – 4)	
Capitalisation	0 (0 – 3)	1 (0 – 3)	0 (0 – 1)	
Spelling	0 (0 – 3)	0 (0 – 1)	0 (0 – 3)	
Lexis	52 (4 – 346)	52 (4 – 338)	56 (10 – 346)	0,669
Omission	5 (0 – 52)	4 (0 – 17)	7 (1 – 52)	
Addition	1 (0 – 58)	2 (0 – 58)	1 (0 – 58)	
Untranslated	40 (0 – 330)	41 (0 – 330)	38 (4 – 307)	
Grammar	14 (2 – 95)	16 (7 – 95)	11 (2 – 44)	0,051
Misselection	9 (0 – 90)	11 (4 – 90)	9 (0 – 32)	
Misordering	4 (0 – 32)	5 (1 – 20)	3 (0 – 32)	
Semantic	12 (1 – 61)	15 (1 – 61)	10 (2 – 26)	0,196
Confusion of senses	3 (0 – 16)	3 (1 – 16)	1 (0 – 8)	
Wrong choice	10 (0 – 53)	9 (0 – 53)	10 (1 – 18)	
Syntax	0 (0 – 3)	0 (0 – 3)	0 (0 – 2)	0,809
Discourse	25 (2 – 104)	**14 (2 – 30)**	**46 (9 – 104)**	**<0,001**
Style	5 (0 – 30)	5 (1 – 24)	6 (0 – 30)	
Variety	7 (0 – 86)	3 (0 – 7)	31 (5 – 86)	
Should not be translated	3 (0 – 19)	3 (0 – 19)	4 (0 – 16)	
Text legibility	1 (0 – 18)	**4 (0 – 18)**	**1 (0 – 8)**	**0,039**
TOTAL	127 (26 – 503)	109 (26 – 478)	148 (36 – 503)	0,224

Table 3. Quantification of each type of error, according to the tool used for translation.

Regarding the analysis carried out by type of study design: case report, systematic review with meta-analysis, observational study or randomized clinical trial, it was found that the second and third types of

study were those that presented a significantly higher frequency of errors' occurrence, of any type, (p=0.003). Errors at the lexical level were also significantly more frequent in these two types of study design (p=0.015). These refer, once again, to segments that remained untranslated, mainly those related to the maintenance of acronyms and abbreviations in English, with at least 14 occurrences of this type being recorded in systematic reviews with meta-analysis. Legibility errors also occurred in a significantly higher proportion in texts relating to observational studies, followed by systematic reviews with meta-analysis (p=0.008).

When the translation results obtained between the different medical specialties were compared, there were no differences in the occurrence of errors in any of the categories.

Table 4 presents an example containing discourse, legibility, and grammatical errors. This example was taken from the systematic review and meta-analysis within the scope of cardiac surgery (Owen *et al.,* 2020, 2737). Through its analysis, it is possible to identify one discourse error, two legibility errors, and one grammatical error in the text obtained from Google Translate, and four discourse errors, two legibility errors, and one grammatical error in the text obtained from ChatGPT.

Let us observe these examples: in the case of the first translation, the discourse error is related to the variety of Portuguese used, that of Brazil. In the second translation, the ChatGPT one, the four errors in this same category have the same origin: "*extracorpórea*" should be replaced by "*extracorporal*," "*pinçamento cruzado*" and "*pinçamento aórtico*" by "*clampagem aórtica*", "*em uma*" by the prepositional contraction "*numa*," and "*liberação*" by "*libertação*." Regarding the first two terms clarified here, "*extracorporal*" and "*clampagem aórtica*" it is recognized that this also concerns a case that could be classified under the style subcategory, since they relate to medical terminology specific to the field of cardiac surgery.

The legibility errors are the same in both target texts and involve the lack of superscript text in the abbreviation that represents the heterogeneity of the results of the meta-analysis to which they refer (*I2* → I^2). Finally, the grammatical error identified relates to both the word order in the Google Translate text (*banda miocárdica da creatina quinase* → *creatina quinase de banda miocárdica*) and the writing of the term by ChatGPT.

Source text	**Target text: Google Translate**	**Target text: ChatGPT**
Methods: … This was followed by a meta-analysis investigating: mortality, crystalloid volume; cardiopulmonary bypass time; cross-clamp time; …; creatine kinase myocardial band (CK-MB) …. Results: … Microplegia used a lower volume of crystalloids and led to a higher spontaneous return of heartbeat, odds ratio (OR) 4.271 (95% confidence intervals [CIs]: 1.935, 9.423; I^2 = 76.57%; P < .001) and a lower requirement for inotropic support, OR: 0.665 (95% CI: 0.47, 0.941; I^2 = 3.53%; P = .021). Microplegia was also associated with a lower CK-MB release ….	Métodos: … Seguiu-se uma meta-análise investigando: mortalidade, volume cristaloide; tempo de circulação extracorporal; tempo de ***pinçamento cruzado***; … **‡banda miocárdica da creatina quinase‡** (CK-MB) …. Resultados: … A microplegia utilizou um menor volume de cristaloides e levou a um maior retorno espontâneo dos batimentos cardíacos, odds ratio (OR) 4,271 (intervalos de confiança [IC] de 95%: 1,935, 9,423; **†I2†** = 76,57%; P < 0,001) e uma menor necessidade de suporte inotrópico, OR: 0,665 (IC 95%: 0,47, 0,941; **†I2†** = 3,53%; P = 0,021). A microplegia também foi associada a uma menor libertação de CK-MB ….	Métodos: … Seguiu-se uma meta-análise investigando: mortalidade, volume de cristaloides, tempo de circulação ***extracorpórea***, tempo de ***pinçamento aórtico***, … **‡creatina quinase banda cardíaca‡** (CK-MB) …. Resultados: … A microplegia utilizou um volume mais baixo de cristaloides e resultou ***em uma*** maior taxa de retorno espontâneo do ritmo cardíaco, razão de chances (OR) 4,271 (intervalos de confiança de 95% [IC]: 1,935, 9,423; **†I2†** = 76,57%; P < .001) e uma menor necessidade de suporte inotrópico, OR: 0,665 (IC 95%: 0,47, 0,941; **†I2†** = 3,53%; P = .021). A microplegia também esteve associada a uma menor ***liberação*** de CK-MB….

Footnotes: * Discourse error; † Legibility error; ‡ Grammatical error.

Table 4. Presentation of a case in which the differences observed between the type of tool used were verified, regarding the occurrence of errors.

Table 5 shows the qualitative analysis, by type of tool used to obtain the translations. All target texts were classified as "understandable", since it was always possible to understand the topic they addressed and the methods the researchers used to achieve their proposed aims. However, all these texts would require human post-editing, in the form of light post-editing, since it would not be necessary to reformulate the entire text, but only to correct errors and less fluent constructions.

Variables	Target Corpus	Google Translate	ChatGPT	P-value
Full text				
Comprehensibility, n (%)	32 (100%)	100%	100%	NA
Fidelity, n (%)				
A lot	6 (18,8%)	3 (18,8%)	3 (18,8%)	0,554
For the most part	22 (68,8%)	10 (62,5%)	12 (75%)	
Complete	4 (12,5)	3 (18,8%)	1 (6,3)	
Fluency				
Non-fluent Portuguese	9 (28,1%)	4 (25%)	5 (31,3%)	0,801
Non-native Portuguese	20 (62,5%)	10 (62,5%)	10 (62,5%)	
Good Portuguese	3 (9,4%)	2 (12,5%)	1 (6,3%)	
Need for post-editing, n (%)	32 (100%)	100%	100%	NA
Need for light post-editing, n (%)	32 (100%)	100%	100%	NA

Footnotes: n – Number of cases; NA – Not applicable.

Table 5. Qualitative assessment of the target texts according to the type of tool.

Table 6 presents some examples taken from the translation of the text relating to the systematic review with meta-analysis, in the area of nephrology, particularly from the methodology and results sections (Fletcher *et al.,* 2022, 5, 7).

Example	Source text	Target text: ChatGPT	Rationale
#1 – Loanwords	The latter area is important, as it could provide **insight** into those symptoms that may be attributable to changes in renal function and may provide potential targets for symptom tracking in CKD.	Esta última área é importante, pois pode fornecer **insights** sobre sintomas que podem ser atribuídos a alterações na função renal e oferecer potenciais alvos para o acompanhamento dos sintomas na DRC.	In this case, it is clear that ChatGPT adopted the loanword that exists in Portuguese, putting it in the plural so that it agrees in number with the rest of the sentence.

#2 - Acronyms	Patients with **CKD** stage 1–5 who were not on RRT were included in 126 (28%) studies.	Pacientes com **CKD** em estágios 1-5 que não estavam em terapia de substituição renal (TSR) foram incluídos em 126 (28%) estudos.	More than just a non-translation of the word, which in this case is an acronym, this is one of the most common acronyms in nephrology: IRC (in Portuguese: **I**nsuficiência **R**enal **C**rónica). Sometimes, algorithms, especially with artificial intelligence, can find the equivalent acronym in their databases, as they did with the acronym "RRT," and even identify the corresponding full term.

Table 6. Examples of the use of foreign words (loanwords) and acronyms.

Discussion

The results of this study demonstrated, overall, the equivalent performance of Google Translate and ChatGPT in obtaining quick translations, after the use of generic interrogative prompts. Despite this similarity in performance, it is possible to discern aspects in which both tools failed and others in which they distinguished themselves. It should be noted that the target texts obtained were, as a whole, understandable and required only light post-editing, according to the human evaluation of their quality. Some of these possible edits are related to issues of fluency and fidelity to the original, but also to the different types of errors found in the translations.

Despite the development of various technologies, the barrier between languages is still evident, especially when it comes to languages other than English, which is widely used not only in scientific but also in academic and technical contexts. Thus, it is virtually impossible to meet all translation needs using only human translators, given the time they require and the associated costs, as well as the lack of reproducibility of the process. Therefore, machine translation and artificial intelligence have been receiving increasing attention from users who need translations in a wide range of areas, including health research (Rivera-Trigueros, 2022, 593, 597).

The quality of the results of machine translation tools is what motivates the improvement of the algorithms underlying these tools.

Thus, several evaluation methods have been proposed and tested, however, the automatic evaluation of these results never seems to be enough, and the contribution of a human evaluator is always needed. There are several forms of evaluation, along with the lack of consensus on which is the most appropriate, hence why a combination of both automatic and human assessments is often suggested, considering the linguistic and scientific complexity of the texts resulting from medical research (Rivera-Trigueros, 2022, 594, 596).

Artificial intelligence is increasingly developing, expanding the variety of application areas and its ability to analyse natural language and to produce natural language-like outputs (Gencer, 2024, 5; Hendy *et al.,* 2023, 1). Some of the advantages of using ChatGPT to obtain machine translations appear to be higher degrees of fluency, largely derived from the use of paraphrases, and the presentation of a less monotonic tone, reflecting to a lesser extent the "*translationese*" sometimes detected with the use of machine translation tools, mainly due to the change in the word order of the source text, something that Google Translate does to a lesser extent, remaining more faithful to the order of the source language. The use of paraphrases could be problematic in the sense that it could imply the significant addition of content that is not present in the source text. Despite this, ChatGPT's fidelity to the source text appears to be adequate and reasonable for the interpretation of the target text, maintaining the original meaning, for the most part, as we have verified throughout this study. However, each study must independently evaluate its results, being suggested a hybrid approach for obtaining translations that uses both machine translation and AI (Hendy *et al.,* 2023, 20-23).

In our study, we observed that ChatGPT is less prone to legibility changes, but when it occurs, it has a direct effect on the way the reader interacts with the text, through the addition or omission of features that are absent in the source text. Other authors have described this addition of features that make the text more fluent and easier to read to a greater extent than machine translation tools. This behaviour may be inappropriate depending on the purpose of the text, as well as the contextual domain in which it belongs (Hendy *et al.,* 2023, 21).

The readability assessment of a text is not free from limitations, namely, the simplicity of the formulas it applies, since these are based on

numerical characteristics of the texts and ignore semantic, cohesion, syntactic or rhetorical issues, which can influence the degree of textual complexity (Collins-Thompson, 2014, 98, 101). Therefore, in the present study, we attempted to overcome this gap by identifying and classifying different types of errors, in order to complement the information provided by the readability scores.

Metrics for studying readability in European Portuguese are limited and base their calculations on counts, excluding textual features that can actively influence the difficulty and complexity of the text. Specifically, in the field of medicine, as in other specialty domains, the information needed to outline new readability evaluation methods is even scarcer, with no properly annotated corpora on this topic, particularly in languages other than English (Antunes, 2019, 12, 13, 17).

Conclusion

In summary, the 32 texts produced by the machine translation and artificial intelligence tools were quantitatively similar, with no significant differences, either in terms of absolute or relative counts of different textual parameters that characterize the texts, or in terms of the readability assessment scores selected. It can therefore be concluded that, at the time, both tools provided results of similar quality regarding the translation of healthcare scientific papers. Qualitatively, the results of comprehensibility, fluency and fidelity to the source text were also similar between the tools used in these texts that are highly complex and specific, as the results corroborate. However, after human review of the produced texts, some gaps were detected in these last two parameters, preventing them from reaching the highest quality.

This study is not without limitations, namely, the size of the corpus selected for this project was limited to 16 source texts and 32 target texts, totalling 48 texts studied. Another limitation is related to the use of readability assessment formulas, both developed for English and adapted for Portuguese, since this mixed approach could influence the results and limit its comparability in this context.

One possible avenue of research involves testing a series of more complex and specific prompts in artificial intelligence tools in order to optimise their translation results. Studying the application of different

interrogative and conversational prompts could be very fruitful in guiding work with ChatGPT.

References

ANTUNES, H., LOPES, T. C., Analyzing the Adequacy of Readability Indicators to a Non-English Language, *Lecture Notes in Computer Science*, 2019, 149-155. https://doi.org/10.1007/978-3-030-28577-7_10

ANTUNES, H. M. M., Automatic Assessment of Health Information Readability. [Master's dissertation, Universidade do Porto], Repositório Aberto da Universidade do Porto, 2019. https://hdl.handle.net/10216/121810.

CHÉRAGUI, M. A., Theoretical Overview of Machine translation, *ICWIT*, 2012, 160-169.

COLEMAN, M., LIAU, L. T., A computer readability formula designed for machine scoring, *JOURNAL OF APPLIED PSYCHOLOGY*, 1975, **60(2)**, 283-284. https://doi.org/10.1037/h0076540

COLINA, S., MARRONE, N., INGRAM, M., SÁNCHEZ, D., Translation Quality Assessment in Health Research: A Functionalist Alternative to Back-Translation, *EVALUATION & THE HEALTH PROFESSIONS*, 2017*,* **40(3)**, 267-293. https://doi.org/10.1177/0163278716648191

COLLINS-THOMPSON, K., Computational assessment of text readability, *RECENT ADVANCES IN AUTOMATIC READABILITY ASSESSMENT AND TEXT SIMPLIFICATION*, 2014, **165(2)**, 97-135. https://doi.org/10.1075/itl.165.2.01col

COSTA, Â., LING, W., LUÍS, T., CORREIA, R., COHEUR, L., A linguistically motivated taxonomy for Machine Translation error analysis, *MACHINE TRANSLATION*, 2015, **29(2)**, 127-161. https://doi.org/10.1007/s10590-015-9169-0

COSTA, P. S., XAVIER, V. M. C., MORAIS, R. S., TIBÉRIO, C. S. B., Estudo Empírico do Grau de Compreensibilidade e Legibilidade dos Relatórios de Administração e da Escala de Locus de Controle nos Cursos de Administração e Ciências Contábeis, *XXX Encontro Nacional Da Anpad,* 2006.

FLETCHER, R. B., DAMERY, S., AIYEGBUSI, L. O., ANDERSON, N., CALVERT, M., COCKWELL, P., FERGUSON, J., HORTON, M., PAAP, S. C. M., SIDEY-GIBBONS, C., SLADE, A., TURNER, N., KYTE, D., Symptom burden and health-related quality of life in chronic kidney disease: A global systematic review and meta-analysis, *PLOS MEDICINE*, 2022, **19(4)**, e1003954. https://doi.org/10.1371/journal.pmed.1003954

GENCER, A., Readability analysis of ChatGPT's responses on lung cancer, *SCIENTIFIC REPORTS*, 2024, **14(1)**. https://doi.org/10.1038/s41598-024-67293-2

GUNNING, R., *The Technique of Clear Writing*, New York, McGraw-Hill, 1952.

HENDY, A., MOHAMED, A., AMR, S., VIKAS, R., MOHAMED, G., HITOKAZU, M., JIN, Y. K., MOHAMED, A., HASSAN, H. A.,How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation, 2023. https://doi.org/10.48550/arXiv.2302.09210

IKRAMOVNA, G. N., LAZIZOVNA, K. D., The problems of translation and some useful strategies in translating english medical terms into uzbek, *SO 'NGI ILMIY TADQIQOTLAR NAZARIYASI*, 2024, **7(4)**, 25-29.

ISMAYILLI, T., Navigating Complexities in Medical Text Translation: Challenges, Strategies, and Solutions, *ACTA GLOBALIS HUMANITATIS ET LINGUARUM*, 2024, **1(2)**, 170-176. https://doi.org/10.69760/aghel.01024080

KAMANDHARI, H. H., The Definitions and the Measurement of Legibility and Readability in Instructional Text Design: An Integrated Literature Review, *INFORMATION, MEDIUM, AND SOCIETY: JOURNAL OF PUBLISHING STUDIES*, 2020, **18(2)**, 1-21. https://doi.org/10.18848/2691-1507/CGP/v18i02/1-21

KENNY, D., Human and machine translation, *MACHINE TRANSLATION FOR EVERYONE: EMPOWERING USERS IN THE AGE OF ARTIFICIAL INTELLIGENCE*, 2022, **18**, 23. https://doi.org/10.5281/zenodo.6759976

KINCAID, J. P., Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, *Chief of Naval Technical Training*, 1975.

MCDONALD, V. S., Accuracy, readability, and acceptability in translation, *Applied Translation,* 2020. https://doi.org/10.51708/apptrans.v14n2.1238

OWEN, M. C., ASOPA, S., SMART, A. N., KING, N., Microplegia in cardiac surgery: Systematic review and meta-analysis, *JOURNAL OF CARDIAC SURGERY*, **2020**, 35(10), 2737-2746. https://doi.org/10.1111/jocs.14895

PEREIRA, S. C., SANTIAGO, M. L., SILVA, R. I., FERREIRA, P., Validation and Cultural Adaptation of the Problem Areas in Diabetes-5 (PAID-5) Scale to European Portuguese, *ACTA MED PORT*, 2023, **37(1)**, 36-41. https://doi.org/10.20344/amp.18995

RIVERA-TRIGUEROS, I., Machine translation systems and quality assessment: a systematic review, *LANGUAGE RESOURCES AND EVALUATION*, 2022, **56(2)**, 593-619. https://doi.org/10.1007/s10579-021-09537-5

SMITH, E. A., SENTER, R. J., *Automated Readability Index*, Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Vol. 66, nº220, 1967.

SMITH, M., TAFFLER, R., Readability and Understandability: Different Measures of the Textual Complexity of Accounting Narrative, *ACCOUNTING, AUDITING & ACCOUNTABILITY JOURNAL*, 1992, **5(4)**. https://doi.org/10.1108/09513579210019549

SUSANTHI, I. G. A. A. D., Translation on Medical Terms into Indonesian, International *JOURNAL OF ARTS AND SOCIAL SCIENCE*, 2023, **6(1)**, 22-29.

Source Corpus References

BRANCH, R. H. K., GATEWOOD, O. M., KUDENCHUK, J. P., MAYNARD, C., SAYRE, R. M., CARLBOM, J. D., EDWARDS, M. R., COUNTS, R. C., PROBSTFIELD, L. J., BRUSEN, R., JOHNSON, N., GUNN, L. M., Diagnostic yield, safety, and outcomes of Head-to-pelvis sudden death CT imaging in post arrest care: The CT FIRST cohort study, *RESUSCITATION*, 2023, **188**, 109785. https://doi.org/10.1016/j.resuscitation.2023.109785

BURNS, C., YEATES, L., SWEETING, J., SEMSARIAN, C., INGLES, J., Evaluating a communication aid for return of genetic results in families with hypertrophic cardiomyopathy: A randomized controlled trial, *JOURNAL OF GENETIC COUNSELLING*, 2023, **32(2)**, 425-434. https://doi.org/10.1002/jgc4.1651

DAVEY, G. M., TEMPERLEY, C. H., O'SULLIVAN, J. N., MARCELINO, V., RYAN, K. O., RYAN, J. É., DONLON, E. N., JOHNSTON, M. S., ROBB, B. W., Minimally Invasive and Open Gastrectomy for Gastric Cancer: A Systematic Review and Network Meta-Analysis of Randomized Clinical Trials, *ANNALS OF SURGICAL ONCOLOGY*, 2023, **30(9)**, 5544-5557. https://doi.org/10.1245/s10434-023-13654-6

FLETCHER, R. B., DAMERY, S., AIYEGBUSI, L. O., ANDERSON, N., CALVERT, M., COCKWELL, P., FERGUSON, J., HORTON, M., PAAP, S. C. M., SIDEY-GIBBONS, C., SLADE, A., TURNER, N., KYTE, D., Symptom burden and health-related quality of life in chronic kidney disease: A global systematic review and meta-analysis, *PLOS MEDICINE,* 2022, **19(4)**, e1003954. https://doi.org /10.1371/journal.pmed.1003954

HALL, R. C., RAMASESHAN, R., REID, A., JONES, D. A., MATHUR, A., Case report: Cytokine therapy and an intracoronary autologous bone marrow-derived cell infusion with Impella support in a patient with dilated cardiomyopathy and a severely reduced ejection fraction, *FRONT CARDIOVASC MED*, 2022, **9**, 1002508. https://doi.org/10.3389/fcvm.2022.1002508

HAYES, N. W., LAING, E., BROWN, R., SILSBY, L., SMITH, L., THOMAS, H., KALOYIROU, F., SHARMA, R., GRIFFITHS, J., HUME-SMITH, H., MARKS, D. S., KESSARIS, N., CHRISTIAN, M., DUDLEY, J., SHENOY, M., MALINA, M., MUORAH, M., WARE, N., YADAV, P., PETERS, J. M., A pragmatic, open-label, randomized controlled trial of Plasma-Lyte-148 versus standard intravenous fluids in children receiving kidney transplants (PLUTO), *KIDNEY INTERNATIONAL*, 2024, **105(2)**, 364-375. https://doi.org/10.1016/j.kint.2023.09.032

HORWOOD, G. A. C., PATEL, D. N., WALKER, D. J., EVANS, R. C., Nonanemic Iron Deficiency in Cardiac Surgery: A Retrospective Observational Study, *JOURNAL OF CARDIOTHORACIC AND VASCULAR ANESTHESIA,* 2024, **38(9)**, 1899-1906. https://doi.org/10.1053/j.jvca.2024.05.039

LAW, M. Y., HSU, C., HINGORANI, R. S., RICHARDS, M., MCMULLAN, M. D., JEFFERIES, H., HIMMELFARB, J., KATZ, R., Randomized controlled trial of remote ischemic preconditioning in children having cardiac surgery, *JOURNAL OF CARDIOTHORACIC SURGERY,* 2024, **19(1)**, 5. https://doi.org/10.1186/s13019-023-02450-8

LAZARUS, B., POLKINGHORNE, K. R., GALLAGHER, M., COGGAN, S., GRAY, N. A., TALAULIKAR, G., KOTWAL, S., INVESTIGATORS, R., Tunneled hemodialysis catheter tip design and risk of catheter dysfunction: an Australian nationwide cohort study, *AMERICAN JOURNAL OF KIDNEY DISEASES,* 2024, **83(4)**, 445-455. https://doi.org/10.1053/j.ajkd.2023.09.021

MCCLAINE, S., FEDOR, J., BARTEL, C., CHEN, L., DURICA, K. C., LOW, C. A., Engagement with daily symptom reporting, passive smartphone sensing, and wearable device data collection during chemotherapy: longitudinal observational study, *JMIR cancer,* 2024, **10(1)**, e57347. https://doi.org/10.2196/57347

NUTTING, C., FINNERAN, L., ROE, J., SYDENHAM, A. M., BEASLEY, M., BHIDE, S., BOON, C., COOK, A., WINTON, D. E., EMSON, M., FORAN, B., FROGLEY, R., PETKAR, I., PETTIT, L., ROONEY, K., ROQUES, T., SRINIVASAN, D., TYLER, J., HALL, E., MACGREGOR, C.. Dysphagia-optimised intensity-modulated radiotherapy versus standard intensity-modulated radiotherapy in patients with head and neck cancer (DARS): a phase 3, multicentre, randomised, controlled trial, *THE LANCET ONCOLOGY,* 2023, **24(8)**, 868-880. https://doi.org/10.1016/S1470-2045(23)00265-6

OWEN, M. C., ASOPA, S., SMART, A. N., KING, N., Microplegia in cardiac surgery: Systematic review and meta-analysis, *JOURNAL OF CARDIAC SURGERY,* 2020, **35(10)**, 2737-2746. https://doi.org/10.1111/jocs.14895

SWEENEY, A. C., QUADER, M., KIM, C., Thrombotic Microangiopathic Anemia After Cardiac Surgery, *JOURNAL OF CARDIOTHORACIC AND VASCULAR ANESTHESIA,* 2024, **38(3)**, 784-787. https://doi.org/10.1053/j.jvca.2023.11.033

POWERS, M. C., VERMA, H., ORLOFF, J., PIONTKOWSKI, J. A., TIERSTEN, A., LAMB, A., GULATI, N., Use of a topical Janus kinase inhibitor in immune checkpoint inhibitor-induced eczematous reaction: a case report, *JOURNAL OF DERMATOLOGICAL TREATMENT,* 2024, **35(1)**, 2336118. https://doi.org /10.1080_/09546634.2024.2336118

THOMAS, A. D., TRAINOR, R., SHEINGOLD, Z., SAMARNEH, M., A Case of COVID-19-Associated Focal Segmental Glomerulosclerosis, *CUREUS*, 2023, **15(4)**. https://doi.org/10.7759/cureus.37547

WINCHESTER, D. E., MERRITT, J., WAHEED, N., NORTON, H., MANJA, V., SHAH, N. R., HELFRICH, C. D., Implementation of appropriate use criteria for cardiology tests and procedures: a systematic review and meta-analysis, *EUROPEAN HEART JOURNAL-QUALITY OF CARE AND CLINICAL OUTCOMES,* 2021, **7(1)**, 34-41. https://doi.org/10.1093/ehjqcco/qcaa029

Raquel MOREIRA is a master's degree student of Arts in Translation and Language Services at the Faculty of Arts and Humanities of the University of Porto, in Portugal. Her previous academic path includes a degree in Cardiopulmonary Physiology and a master's degree in Cardiovascular Pathophysiology. The author has worked under a fellowship program at the Faculty of Medicine of the University of Porto and now develops her

research, within her new master's degree, in the field of medical translation due to her interest in these interdisciplinary fields of science.